Towards Establishing a Standardized Magnetic Resonance Imaging Scoring System for Temporomandibular Joints in Juvenile Idiopathic Arthritis

Tolend, Mirkamal A ; Twilt, Marinka ; Cron, Randy Q ; Tzaribachev, Nikolay ; Guleria, Saurabh ; von Kalle, Thekla ; Koos, Bernd ; Miller, Elka ; Stimec, Jennifer ; Vaid, Yoginder ; Larheim, Tore A ; Herlin, Troels ; Spiegel, Lynn ; Inarejos Clemente, Emilio J ; Moineddin, Rahim ; van Rossum, Marion A ; Saurenmann, Rotraud K ; Doria, Andrea S ; Kellenberger, Christian J

Abstract: OBJECTIVES: The temporomandibular joints (TMJs) are frequently affected in children with juvenile idiopathic arthritis (JIA). Early detection is challenging, as major variation is present in scoring TMJ pathology on Magnetic Resonance Imaging (MRI). Consensus-driven development and validation of a MRI scoring system for TMJs has important clinical utility in timely improvement of diagnosis, and serving as an outcome measure. We report on a multi-institutional collaboration towards developing a TMJ MRI scoring system for JIA. METHODS: Seven readers independently assessed MRI scans from 21 patients (42 TMJs, age range 6-16y) using three existing MRI scoring systems from American, German, and Swiss institutions. Reliability scores, scoring system definitions and items were discussed among 10 JIA experts through two rounds of Delphi surveys, nominal group voting, and subsequent consensus meetings to create a novel TMJ MRI scoring system. RESULTS: Average-measure intraclass correlation coefficients (avICC) for the total scores of all three scoring systems were highly reliable at 0.96 each. Osteochondral items showed higher reliability than inflammatory items. An additive system was deemed preferable for assessing minor joint changes over time. Eight items were considered sufficiently reliable and/or important for integration into the consensus scoring system: bone marrow edema and enhancement (avICC=0.57-0.61; %SDD=±45-63% prior to re-defining), condylar flattening (0.95-0.96; ±23-28%), effusions (0.85-0.88; ±25-26%), erosions (0.94; ±20%), synovial enhancement and thickening (previously combined; 0.90-0.91; ±33%), and disk abnormalities (0.90; ±19%). CONCLUSION: A novel TMJ MRI scoring system was developed by consensus. Further iterative refinements and reliability testing are warranted in upcoming studies.

DOI: https://doi.org/10.1002/acr.23340

Rotraud K; Doria, Andrea S; Kellenberger, Christian J (2018). Towards Establishing a Standardized Magnetic Resonance Imaging Scoring System for Temporomandibular Joints in Juvenile Idiopathic Arthritis. Arthritis Care Research, 70(5):758-767.
DOI: https://doi.org/10.1002/acr.23340

Article type : Original Article

Title: Towards Establishing a Standardized Magnetic Resonance Imaging Scoring System for Temporomandibular Joints in Juvenile Idiopathic Arthritis

Running Head: TMJ MRI Scoring System for JIA

Authors: Mirkamal A. Tolend, Marinka Twilt, Randy Q. Cron, Nikolay Tzaribachev, Saurabh Guleria, Thekla von Kalle, Bernd Koos, Elka Miller, Jennifer Stimec, Yoginder Vaid, Tore A. Larheim, Troels Herlin, Lynn Spiegel, Emilio Inarejos, Rahim Moineddin, Marion A. van Rossum, Rotraud K. Saurenmann, Andrea S. Doria*, Christian J. Kellenberger*

*Co-senior authors

Author affiliations (please use this instead of the manuscript portal information):

1st author: Mirkamal A. Tolend, Department of Diagnostic Imaging, The Hospital for Sick Children, Toronto, Canada.

2nd author: Marinka Twilt, MD, PhD, Division of Rheumatology, Department of Paediatrics, Alberta Children's Hospital, Calgary, Canada.

3rd author: Randy Q. Cron, MD, PhD, Department of Pediatrics, Children's Hospital of Alabama, Birmingham, Alabama, USA.

4rd author: Nikolay Tzaribachev, MD, Pediatric Rheumatology Research Institute, Bad Bramstedt, Germany.

5th author: Saurabh Guleria, MD, Austin Radiological Association, Austin, Texas, USA.

6th author: Thekla von Kalle, MD, Department of Pediatric Radiology, Radiologisches Institut, Olgahospital Klinikum Stuttgart, Stuttgart, Germany.

7th author: Bernd Koos, DMD, Department of Orthodontics, University Hospital Tübingen, Tübingen, Germany.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/acr.23340

8th author: Elka Miller, MD, Department of Medical Imaging, Children's Hospital of Eastern Ontario, Ottawa, Canada.

9th author: Jennifer Stimec, MD, Department of Diagnostic Imaging, The Hospital for Sick Children, Toronto, Canada.

10th author: Yoginder Vaid, MD, Department of Radiology, Children's Hospital of Alabama, Birmingham, Alabama, USA.

11th author: Tore A. Larheim, DDS, PhD, Department of Maxillofacial Radiology, Institute of Clinical Dentistry, University of Oslo, Blindern, Oslo, Norway.

12th author: Troels Herlin, MD, Department of Pediatrics, Aarhus University Hospital, Aarhus, Denmark.

13th author: Lynn Spiegel, MD, Department of Rheumatology, The Hospital for Sick Children, Toronto, Canada.

14th author: Emilio Inarejos, MD, Department of Diagnostic Imaging, Hospital Sant Joan de Deu, Barcelona, Spain.

15th author: Rahim Moineddin, PhD, Department of Family and Community Medicine, University of Toronto, Toronto, Canada.

16th author: Marion A. van Rossum, MD, PhD, Department of Pediatrics, Emma Children's Hospital, Academic Medical Centre, and Department of Pediatric Rheumatology, Amsterdam Rheumatology and Immunology Center, Reade, Amsterdam, The Netherlands.

17th author: Rotraud K. Saurenmann, MD, Division of Rheumatology, University Children's Hospital, Zürich, Switzerland.

18th author (co-senior author): Andrea S. Doria, MD, PhD, MSc, Department of Diagnostic Imaging, The Hospital for Sick Children, Toronto, Canada.

19th author (co-senior author): Christian J. Kellenberger, MD, Department of Diagnostic Imaging, University Children's Hospital, Zürich, Switzerland.

Corresponding Author Contact:

Andrea S. Doria, MD, PhD, MSc
Professor, Associate Vice-Chair of Research (Injury, Repair and Inflammation),
Department of Medical Imaging, University of Toronto
Radiologist, Senior Scientist, Research Director,
Department of Diagnostic Imaging, The Hospital for Sick Children
555 University Avenue, 2nd floor, Toronto, ON M5G1X8
Phone: 416-813-6079, Fax: 416-813-7591
Email: andrea.doria@sickkids.ca

Funding Information:

Mirkamal Tolend was supported by graduate student stipend awards from the Hospital for Sick Children (Toronto, Canada), and the Queen Elizabeth II/Edward Dunlop Foundation (Canada). Meeting expenses of the group were partly funded by a knowledge translation grant from the Canadian Institute of Health Research. Funding agencies were not involved in the study design or the analysis, interpretation, or reporting of the results. The authors do not have any conflicts of interest.

Abstract:

Objectives. The temporomandibular joints (TMJs) are frequently affected in children with juvenile idiopathic arthritis (JIA). Early detection is challenging, as major variation is present in scoring TMJ pathology on Magnetic Resonance Imaging (MRI). Consensus-driven development and validation of a MRI scoring system for TMJs has important clinical utility in timely improvement of diagnosis, and serving as an outcome measure. We report on a multi-institutional collaboration towards developing a TMJ MRI scoring system for JIA.

Methods. Seven readers independently assessed MRI scans from 21 patients (42 TMJs, age range 6-16y) using three existing MRI scoring systems from American, German, and Swiss institutions. Reliability scores, scoring system definitions and items were discussed among 10 JIA experts through two rounds of Delphi surveys, nominal group voting, and subsequent consensus meetings to create a novel TMJ MRI scoring system.

Results. Average-measure intraclass correlation coefficients (avICC) for the total scores of all three scoring systems were highly reliable at 0.96 each. Osteochondral items showed higher reliability than inflammatory items. An additive system was deemed preferable for assessing minor joint changes over time. Eight items were considered sufficiently reliable and/or important for integration into the consensus scoring system: bone marrow edema and enhancement (avICC=0.57-0.61; %SDD=±45-63% prior to re-defining), condylar flattening

(0.95-0.96; ±23-28%), effusions (0.85-0.88; ±25-26%), erosions (0.94; ±20%), synovial enhancement and thickening (previously combined; 0.90-0.91; ±33%), and disk abnormalities (0.90; ±19%).

Conclusion. A novel TMJ MRI scoring system was developed by consensus. Further iterative refinements and reliability testing are warranted in upcoming studies.

Significance and Innovations:

Significance:

- Contrast-enhanced MRI remains the gold standard for detecting early arthritic changes in the TMJ; however, image interpretation is subjective and needs standardization.
- Multi-institutional, consensus-driven development and validation of a scoring system for interpreting MRI of TMJs is essential for standardizing outcome assessment in clinical studies and for interval monitoring of JIA disease activity.

Innovations:

- Three independently developed TMJ MRI scoring systems for JIA were compared and unified into a new scoring system by consensus.
- Inter-reader reliability of MRI findings among seven readers from different institutions was assessed in addition to Delphi surveys and nominal group technique to determine the item selection, definition, and grading for the newly developed scoring system.

Manuscript Text:

Juvenile idiopathic arthritis (JIA) is the most frequent cause of chronic inflammatory arthritis in childhood, with a prevalence of approximately 1 in 1,000 children (1). Inflammation, structural changes or joint damage in the temporomandibular joint (TMJ) observable with magnetic resonance imaging (MRI) may occur in 40-90% of patients with JIA, depending on the subpopulation examined and the interpretation of the images (2). However, these changes often develop silently without symptoms, and irreversible facial deformities and functional impairments can be present by the time clinical symptoms appear (3). Therefore, it is essential to identify TMJ arthritis early. Hopefully, early TMJ diagnosis may delay, stop or even reverse the course of the disease when adequately treated. Contrast-enhanced MRI is currently the diagnostic standard, since it allows the visualization of inflammatory changes, structural abnormalities, and damage accrual in the TMJ, which cannot be reliably assessed by physical examination or other diagnostic tests (4–8). However, the specific use of MRI for TMJ arthritis is not standardized at present. There is a recognized need to organize a consensus effort to determine which anatomical and pathological features to assess on MRI, and how to grade the changes in these items to best reflect an MRI-based indication of disease severity. Generating such a model for imaging-based assessment of TMJ arthritis in JIA requires a formalized consensus approach, particularly because there is no reliable clinical outcome measure against which to correlate MRI observations. Nevertheless, this is theoretically feasible, as similar efforts are being coordinated with other groups within the MRI in JIA special interest group under the Outcome Measures in Rheumatoid Arthritis and Clinical Trials (OMERACT) umbrella to develop respective MRI grading systems in large and small joints. Validating and translating such an imaging-based outcome measure into clinical practice and research, when used together with functional clinical orofacial outcome

measure of TMJ arthritis (9), will have tremendous utility for standardizing outcome reporting in research, as well as interval monitoring of disease status in JIA patients to inform treatment selection.

The aims of this study were to assess the reliability of three existing MRI scoring systems for assessment of arthritis in the TMJs of JIA patients, and to establish standardized consensus guidelines for interpretation of MRI examinations of TMJs under the auspices of the OMERACT MRI in JIA special interest group. To accomplish this, first, the inter-reader reliability of three existing TMJ MRI scoring systems for JIA was evaluated. Then, considering these reliability results and the various methods for scoring these TMJ features, semi-structured consensus-forming techniques were used to develop a single TMJ MRI scoring system with refined item definitions and grading methods.

Materials & Methods

Reliability Exercise

Three TMJ MRI Scoring Systems

Three existing scoring TMJ MRI systems were identified from three academic institutions for assessment of inter-reader reliability. The German (10) and American (11) systems were semi-quantitative and additive, where a total score is determined by the sum of all individually graded constituent items. The German scoring system (10) contained five items: synovial fluid, bone marrow edema, synovitis, erosions, and changes in condylar shape. (Appendix 1). All five items were graded from 0 to 3, with equal measurement spacing across the grades. The American scoring system (11)

contained seven items with variable item weights, ranging from 0-1 (marrow edema, osteophyte formation), 0-3 (effusion, synovial enhancement, disk abnormalities, erosions, and pannus), or 0-4 (condyle abnormalities) (Appendix 2). The Swiss system (12) was a progressive system, where sub-scores are determined 0-4 based on the most severe feature (Appendix 3).

Readers and MRI Sample

Seven investigators participated in the TMJ MRI reading exercise, which consisted of independently scoring the images using the three existing MRI scoring systems. To improve the representativeness of the reader sample, and hence the external validity of this inter-reader reliability exercise (13), these seven readers were selected from seven different centers in four countries (Appendix 4). The reader sample included five board-certified pediatric radiologists, one pediatric rheumatologist and one orthodontist, all of whom were fellowship-trained in their respective specialties and with varying experience reading TMJ MRI (5-20 years post fellowship training). The readers attended a video tutorial session using five non-study exams to standardize their use of the scoring systems.

Sample size was determined to detect a difference in intraclass correlation coefficients between 0.6 and 0.8, at a two-tailed 5% error level and with 80% power. For seven readers, approximately 20 independent measurements would be needed to achieve these parameters (14). Twenty-one bilateral TMJ MRI studies from 2006 through 2010 of boys and girls (age <18) with clinically diagnosed or suspected JIA were selected from a single site (University Children's Hospital Zurich, Switzerland). The cases were sampled purposively by one author (C.J.K.) to represent a variety of MRI-based inflammatory and osteochondral changes. All MRI scans were performed using 1.5 Tesla (Signa MR/i Twinspeed, GE Medical Systems, Milwaukee, WI, USA) with a dedicated TMJ coil in the closed mouth position (8). Detailed description of the standardized TMJ MRI protocol used for these cases can be found in Appendix 5. The study was reviewed and approved by the Hospital for Sick

Children (Toronto, Canada) Research Ethics Board (application number 1000042164), and was conducted in accordance with local health research regulations.

Statistical Analysis

The inter-reader reliability of the overall and constituent item scores of each of the MRI scoring systems were calculated, using the 2-way random, single- and average-measure absolute agreement intraclass correlation coefficients (sICC, avICC), smallest detectable difference (15) (%SDD, adjusted as percent of highest score), and percent exact and close agreement (PEA, PCA). Arbitrarily, average-measure intraclass correlation coefficients (avICC) with lower-bound 95% confidence interval at>0.8 and smallest detectable difference (%SDD)<30% of maximum grade were considered sufficiently reliable, similar to a previous study with six readers (16). Sensitivity analysis of the reliability coefficients was performed excluding the observations from two readers with missing data (Appendix 6). SAS® version 9.4 (SAS Institute Inc, Cary, NC, USA) was used for analyses.

Scoring System Development - Consensus Meetings

Structured communication and consensus-forming techniques (17), including Delphi and nominal group technique (NGT) surveys, were utilized in developing the consensus scoring system to achieve informed and unbiased member participation. A detailed flowchart visualizing the overall process is available in Appendix 7, also described below.

Two rounds of Delphi surveys were completed to generate consensus for item definitions, grading, and MRI imaging sequence recommendations. The MRI reading exercise participants were asked to provide detailed feedback on issues regarding appropriateness and ease-of-use of item definitions

and grading methods in the three existing systems. The first Delphi survey was developed by summarizing these suggestions. The survey was electronically distributed to members of the OMERACT MRI in JIA special interest group, consisting of pediatric and maxillofacial radiologists, pediatric rheumatologists, and orthodontists. Ten members within this group with expertise and interest in TMJ imaging participated in the survey and subsequent consensus meetings. Descriptive statistics on the first Delphi survey, as well as the additional suggestions and new questions raised by respondents were aggregated to develop the second Delphi survey. This second survey was completed at the face-to-face consensus meeting, where the expert panel discussed the various options for item definition and grading in consideration of the first Delphi survey responses and reliability scores, then reached a consensus decision by voting a second time iteratively on the same set of questions (as Delphi Survey 2), with ≥80% agreement considered satisfactory.

Item selection for the scoring system was performed by NGT at the consensus meeting, between the first and second Delphi surveys (Appendix 7). A template set of scoring system items was generated from the three systems to develop the first NGT survey. Considering the reliability reading exercise results, and the item generation and selection related questions from the first Delphi survey, the same 10 experts voted anonymously on these items for inclusion into the scoring system. The results of the first NGT were presented to the participating panel of experts at the consensus meeting along with the item-wise reliability scores from the case reading exercise, followed by a discussion session to clarify and/or modify item options. The panel then anonymously voted on the items again in the second NGT survey, with further discussions of the results. One final open voting (≥80% agreement) was conducted to select the set of items for the first iteration of the consensus TMJ MRI scoring system. Subsequent unstructured consensus meetings were held with a broader group of experts (n=17) between 2015-2016 to finalize definitions and measurement cutoffs.

Results

Reliability Exercise

Clinical characteristics of the 21 patients selected for the reliability exercise are summarized in Table 1. As is typical with JIA in general, these patients were predominantly female and most presented with oligoarticular and polyarticular JIA. On physical examination, more than half of the patients had evidence of growth changes in the jaw, including micrognathia and asymmetry. Similar to other reports of frequently asymptomatic TMJ arthritis in JIA (5,6,18), only a quarter of the patients (n=5/21) reported TMJ pain, while over half (n=11/21) showed MRI evidence of inflammatory disease with at least 25% of total domain score (seven reader average) according to the German and Swiss scoring systems, and nine out of 21 in the American system. Three-quarters of the children were actively receiving a variety of systemic anti-inflammatory treatments at the time of the TMJ MRI.

The reliability exercise showed high inter-reader agreement in total scores for each of the three scoring systems. The avICCs for the American, German, and Swiss systems were each 0.96, corroborated by excellent %SDD scores (Table 2). When the items were grouped into either inflammatory or damage domains, all three scoring systems showed increased reliability in the damage domain items in comparison to active inflammatory changes. Changes more easily visualized on MRI images, such as the contour and integrity of bony regions, showed higher reliability than changes in soft tissues in all three scoring systems. Reliability of equivalent items that appear in both of the additive systems, specifically bone marrow edema, joint effusion, erosions, and condylar flattening, did not differ between the two scoring systems, despite minor differences in item definitions and measurement cutoffs (Appendix 1-3). Several items, including bone marrow edema, pannus, osteophytes and effusion, failed to meet the a priori thresholds for reliability, i.e., lower-

bound 95% confidence interval of avICC at 0.8 or above, and/or %SDD at 30% or below (Table 2). The ICC coefficients were not affected when recalculated excluding two readers with missing data (Appendix 6).

Consensus Item Selection

From the additive scoring systems, five items were considered sufficiently reliable and/or important for assessment: bone marrow edema, condylar flattening, effusion, erosions, and synovial thickening. The items synovial enhancement and pannus (from the American system) and their equivalent combined item "synovitis" from the German system showed similar but relatively moderate reliability (Table 2). Two additional items, disk abnormalities, and bone marrow enhancement (which was previously combined with bone marrow edema) were agreed to be further tested as ancillary items in addition to the six core items. Since the main objective is to make the consensus scoring system more responsive to early changes, the members agreed to use an additive scoring approach to improve discriminatory ability. Differentiating between moderate and severe levels of grading was believed to be inconsequential for clinical management decisions, hence these levels were collapsed to improve reliability and ease-of-use. Definitions and grading thresholds as determined from the consensus discussions are presented in Table 3.

Scoring System Construction

Types of challenges identified and addressed by consensus included unspecific item definitions and measurement criteria, practical challenges in measurements, disagreements in grading thresholds, and item importance. These are described below.

Synovial Changes

The presence of active synovitis is paramount in the evaluation of disease activity in JIA. Synovial thickening and synovial enhancement were agreed to be graded separately, since these two items are assessed independently in different MRI sequences. Moreover, grading these separately would allow for intermediate gradations in synovial changes. For example, non-enhancing synovium may represent residual prior disease which is not actively inflamed. It may not be appropriate to measure synovial enhancement by thickness, or synovial thickening in post-contrast images, since in small joints such as the wrist and the TMJ, the contrast material diffuses into the joint space from the synovium almost immediately after injection, making the two items indistinguishable. Early enhancement of joint fluid in the TMJ has been observed in multiple studies (19–21), hence this limitation led to renaming synovial enhancement to joint enhancement. To conserve independence between joint enhancement and synovial thickening measurements, synovial thickening was agreed to be measured by thickness on T2-weighted images, and joint enhancement by the spread or extension of enhancing region on T1-weighted images, acquired immediately following intravenous injection of Gadolinium based contrast agent (Figure 1). Synovial thickening was decided to be measured at its thickest point in pre-contrast sequences. Focal thickness of >1 and ≤2mm would be graded as mild, or moderate/severe if the thickness is focally >2mm (Table 3). Presence of enhancement was to be measured by hyperintense signal (isointense to blood vessels) in areas exceeding those which correspond to joint effusion (as confirmed on T2-weighted sequence), and graded as local (mild) or diffusely involving the entire joint (moderate/severe) (12).

Osteochondral Changes and Joint Effusion

Areas for improvement were identified also for items with relatively high inter-reader reliability, including condylar flattening and erosions, disk abnormalities, and joint effusion. Grading condylar flattening by the “loss of condyle height” was thought to be unreliable, since a baseline, rounded

condyle morphology had to be assumed and imagined by the reader for grading the amount of flattening. Instead, a grading system by the length of the flattened area was proposed. Since the size of the condyle increases with age, the length should be graded by the proportion of condylar surface flattened. The use of percentage cutoffs was considered, but to improve measurement reliability in this small joint, grading by "part" or "whole width" of the condyle was adopted (Figure 2 vs 1). Similarly, irregularity on the condylar surface was agreed to be graded as part or whole width of condyle, and the presence of deep cortical breaks would be an additional demarcating criterion between mild and moderate/severe levels (Table 3). Severe fluid effusion in the joint spaces was previously defined with different cutoffs for thickness of the fluid layer, such as >1mm (when diffuse, ordinal measure) in the American and >4mm (diffuse, interval measure) in the German scoring systems. By consensus, the cutoff for moderate/severe effusion was changed to >2mm, similar to synovial thickening (Table 3). Both the effusion and synovial thickening were decided to be measured on the same T2-weighted fat-saturated sequence, since a difference in intensity of the signal could differentiate these two features (Figure 1). Disk abnormalities were included in the consensus system as an ancillary item for further investigation of its predictive value.

Bone Marrow Changes

Reference regions to compare the bone marrow signal were discussed. Despite the low reliability in measuring this item and its uncertain functional significance, at this early stage in the scoring system's development and testing, the consensus was that it should be included in the scoring system, but scored separately as bone marrow edema and bone marrow enhancement. Defining and grading these two items separately in different sequences would allow for the assessment of their correlation as well as their individual reliability. This would potentially lead to a better understanding of their significance, and help refine the measurement definitions and imaging protocol in

subsequent iterations. For measurement, the group decided that the signal intensity of the condylar bone marrow should be compared with the contralateral TMJ in unilateral disease involvement, and the bone marrow of the mandibular ramus in bilateral disease. It was assumed that the hematopoietic bone marrow, if present, would be similar in the mandibular ramus compared to the condyle. These two items were reduced to binary grading, in response to concerns that the TMJ condyle is very small in children, making 3 or 4-level ordinal measurements unreliable.

Discussion

Although scoring of TMJs by MRI in patients with JIA has been explored for monitoring of disease course (22,23), this study is the first to compare three different TMJ scoring systems, and to, through data-driven and semi-structured consensus techniques, define the items and grading for interpretation of TMJ MRI examinations.

Single-institution reliability studies may show higher reliability among the raters compared to multi-institution studies, partly due to the shared training among readers. In this study, we have minimized this source of non-independence by studying readers from multiple institutions, professions, and levels of expertise, which should make our assessment of reliability more representative. Furthermore, the right-skewed distribution of disease severity in our sampled cohort could have further reduced statistical efficiency and lowered the correlation coefficients. However, using a sample with mostly mild-moderate level of MRI-observable changes, as we have done, may better approximate the scoring system's real-world performance. Therefore, reliability values from the literature may be incomparable with the current exercise. Instead, using a constant, yet representative reader and image samples across the three scoring systems allowed for controlled assessment of various item definition and grading methods. Overall, the items to be measured,

rather than the particular method of scoring these items, showed a greater impact on inter-reader reliability. The poor agreement observed for measuring the item bone marrow edema is consistent with a previous seven-reader TMJ MRI reading study (24), which also showed relatively higher agreement for joint effusion and disk abnormalities. Specifying anatomical regions for comparing the bone marrow signal, as well as further work into the normal maturation of the bone marrow as seen by MRI will be helpful in improving the reliability of this item.

While the three existing MRI scoring systems were sufficiently reliable for evaluating most soft-tissue and osteochondral changes in TMJs of JIA patients, our consensus-driven discussions favored an additive system, with improvements to certain item definitions and grading criteria for achieving a high level of responsiveness to enhance the consensus scoring system's utility as a longitudinal outcome measure. Thus, a single TMJ MRI scoring system for assessment of both inflammatory and damage-related TMJ changes separately in children with JIA was established using the best features of the previously published approaches. Additionally, the development of a standardized, consensus-developed TMJ MRI protocol will help reduce inter-site differences in TMJ MRI. Subsequent studies are planned to test this new TMJ scoring system's feasibility, reliability, validity, and responsiveness. The use of a formalized, inclusive, and consensus-based development process, such as the one reported in this paper, is a recommended method for improving the feasibility and endorsement of outcome measures. To improve ease-of-use and reliability, our group will subsequently focus on testing and improving the specification of the definitions and grading criteria, constructing a reference TMJ MRI grading atlas to illustrate the variations in scored items, and conducting quantitative studies for improving differentiation of normal and mildly abnormal TMJ findings. Validation testing may be difficult when contrast-enhanced MRI is already the most informative method for the assessment of TMJ. However, cross-sectional and longitudinal validation can still be rigorous, using multiple-hypothesis testing designs based on best available understanding of JIA in TMJ.

Due to the small sample size, our study was neither designed nor powered to explore the correlation of clinical characteristics with scores obtained from the MRI scoring systems. However, larger studies have observed the absence of clinical correlations with findings on MRI, stressing the utility of MRI in the early detection of TMJ involvement to prevent irreversible orofacial changes (6,8). Differences in the number of grading categories between the scoring systems, and the presence of missing data from two readers made it more appropriate to use ICC coefficients instead of kappa and percent agreement. Intra-reader reliability was not assessed in this exercise, though it is generally equal to or higher than inter-reader reliability. Nevertheless, our multi-institution exercise was helpful for comparing the relative reliability of items and scoring methods, and identifying areas for improvement.

Conclusions

Consensus-lead, multi-institutional evaluation found three existing TMJ MRI scoring systems to be sufficiently reliable, although challenges were identified that guided the construction of a new consensus scoring system. In response to the need for a reliable and responsive outcome measure for assessing minor TMJ changes over time, an additive TMJ MRI scoring system optimized for early arthritic changes was developed by multi-institutional consensus. Further testing and iterative refinements will be conducted in upcoming studies from our OMERACT TMJ MRI in JIA special interest group to assess and improve this novel MRI-based outcome measure for use in medical practice and clinical trials in children with JIA and TMJ arthritis.

Acknowledgement

The authors thank Dr. Tal Laor for her contributions to the scale development process in its early stages. Preliminary results on reliability and discussions related to the scoring system development were presented in electronic poster format at the 2015 European Society for Pediatric Radiology meeting.

References

1. Manners PJ, Bower C. Worldwide prevalence of juvenile arthritis why does it vary so much? *J Rheumatol* 2002;29:1520–30.

2. Larheim TA, Doria AS, Kirkhus E, Parra DA, Kellenberger CJ, Arvidsson LZ. TMJ imaging in JIA patients—An overview. *Semin Orthod* 2015;21:102–10.

3. Twilt M, Mobers SMLM, Arends LR, Cate R ten, Suijlekom-Smit L van. Temporomandibular involvement in juvenile idiopathic arthritis. *J Rheumatol* 2004;31:1418–22.

4. Munir S, Patil K, Miller E, Uleryk E, Twilt M, Spiegel L, et al. Juvenile idiopathic arthritis of the axial joints: a systematic review of the diagnostic accuracy and predictive value of conventional MRI. *AJR Am J Roentgenol* 2014;202:199–210.

5. Weiss PF, Arabshahi B, Johnson A, Bilaniuk LT, Zarnow D, Cahill AM, et al. High prevalence of temporomandibular joint arthritis at disease onset in children with juvenile idiopathic arthritis, as detected by magnetic resonance imaging but not by ultrasound. *Arthritis Rheum* 2008;58:1189–96.

6. Koos B, Twilt M, Kyank U, Fischer-Brandies H, Gassling V, Tzaribachev N. Reliability of clinical symptoms in diagnosing temporomandibular joint arthritis in juvenile idiopathic arthritis. *J Rheumatol* 2014;41:1871–7.

7. Colebatch-Bourn AN, Edwards CJ, Collado P, D'Agostino M-A, Hemke R, Jousse-Joulin S, et al. EULAR-PReS points to consider for the use of imaging in the diagnosis and management of juvenile idiopathic arthritis in clinical practice. *Ann Rheum Dis* 2015;74:1946–57.

8. Muller L, Kellenberger CJ, Cannizzaro E, Ettlin D, Schraner T, Bolt IB, et al. Early diagnosis of temporomandibular joint involvement in juvenile idiopathic arthritis: a pilot study comparing clinical examination and ultrasound to magnetic resonance imaging. *Rheumatology* 2009;48:680–5.

9. Stoustrup P, Twilt M, Spiegel L, Kristensen KD, Koos B, Pedersen TK, et al. Clinical Orofacial Examination in Juvenile Idiopathic Arthritis: International Consensus-based Recommendations for Monitoring Patients in Clinical Practice and Research Studies. *J Rheumatol* 2017;44:326–33.

10. Koos DB, Tzaribachev N, Bott S, Ciesielski R, Godt A. Classification of temporomandibular joint erosion, arthritis, and inflammation in patients with juvenile idiopathic arthritis. *J Orofac Orthop Fortschritte Kieferorthopädie* 2013;74:506–19.

11. Vaid YN, Dunnavant FD, Royal SA, Beukelman T, Stoll ML, Cron RQ. Imaging of the temporomandibular joint in juvenile idiopathic arthritis. *Arthritis Care Res* 2014;66:47–54.

12. Kellenberger CJ, Arvidsson LZ, Larheim TA. Magnetic resonance imaging of temporomandibular joints in juvenile idiopathic arthritis. *Semin Orthod* 2015;21:111–20.

13. Obuchowski NA, Zepp RC. Simple steps for improving multiple-reader studies in radiology. *Am J Roentgenol* 1996;166:517–21.

14. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med* 1998;17:101–10.

15. Lassere M, Boers M, Heijde D van der, Boonen A, Edmonds J, Saudan A, et al. Smallest detectable difference in radiological progression. *J Rheumatol* 1999;26:731–9.

16. Lassere M, McQueen F, Østergaard M, Conaghan P, Shnier R, Peterfy C, et al. OMERACT Rheumatoid Arthritis Magnetic Resonance Imaging Studies. Exercise 3: an international multicenter reliability study using the RA-MRI Score. *J Rheumatol* 2003;30:1366–75.

17. Ruperto N, Meiorin S, Iusan S, Ravelli A, Pistorio A, Martini A. Consensus procedures and their role in pediatric rheumatology. *Curr Rheumatol Rep* 2008;10:142–6.

18. Ringold S, Torgerson TR, Egbert MA, Wallace CA. Intraarticular Corticosteroid Injections of the Temporomandibular Joint in Juvenile Idiopathic Arthritis. *J Rheumatol* 2008;35:1157–64.

19. Smith HJ, Larheim TA, Aspestrand F. Rheumatic and nonrheumatic disease in the temporomandibular joint: gadolinium-enhanced MR imaging. *Radiology* 1992;185:229–34.

20. Kottke R, Saurenmann RK, Schneider MM, Müller L, Grotzer MA, Kellenberger CJ. Contrast-enhanced MRI of the temporomandibular joint: findings in children without juvenile idiopathic arthritis. *Acta Radiol* 2015;56:1145–52.

21. Kalle T von, Winkler P, Stuber T. Contrast-enhanced MRI of normal temporomandibular joints in children--is there enhancement or not? *Rheumatol Oxf Engl* 2013;52:363–7.

22. Stoll ML, Vaid YN, Guleria S, Beukelman T, Waite PD, Cron RQ. Magnetic Resonance Imaging Findings following Intraarticular Infliximab Therapy for Refractory Temporomandibular Joint Arthritis among Children with Juvenile Idiopathic Arthritis. *J Rheumatol* 2015;42:2155–9.

23. Lochbühler N, Saurenmann RK, Müller L, Kellenberger CJ. Magnetic Resonance Imaging Assessment of Temporomandibular Joint Involvement and Mandibular Growth Following Corticosteroid Injection in Juvenile Idiopathic Arthritis. *J Rheumatol* 2015;42:1514–22.

24. Takano Y, Honda K, Kashima M, Yotsui Y, Igarashi C, Petersson A. Magnetic resonance imaging of the temporomandibular joint: a study of inter- and intraobserver agreement. *Oral Radiol* 2004;20:62–7.

Legends of Tables and Figures

Table 1: Clinical characteristics of the cohort of 21 patients whose MRI examinations of temporomandibular joints were used for the reliability exercise. Laboratory and physical examination test values are those available at the closest date within three months in relation to the study MRI date. Abbreviations: ANA, Anti-Nuclear Antibody; DMARD, Disease Modifying Anti-Rheumatic Drug; HLA-B27, Human Leukocyte Antigen B27; JIA, Juvenile Idiopathic Arthritis; RF, Rheumatoid Factor; SD, Standard Deviation; TMJ, Temporomandibular Joint.

Table 2: Inter-reader reliability scores of the three MRI scoring systems and their constituent items. AvICC >0.8 and SDD% <30 were the arbitrary thresholds for acceptable reliability (16). *Since these three items are only scored in the American system, they were excluded from the analysis, which resulted in the creation of a new subgroup (American equivalent). This method enabled comparability between the three systems. Abbreviations: avICC, average-measure Intraclass Correlation Coefficient; CI, Confidence Interval; PCA, Percent Close Agreement; PEA, Percent Exact Agreement; SDD, Smallest Detectable Difference; sICC, single-measure Intraclass Correlation Coefficient.

Table 3 A, B): The consensus MRI scoring system for temporomandibular joint involvement in juvenile idiopathic arthritis. The items are grouped by A) Inflammatory and B) Damage domains. Each joint is scored independently.

Figure 1: Seven-year-old girl with polyarticular JIA. Corresponding sagittal oblique, fat-saturated T2-weighted image pre-contrast (**a**) and fat-saturated T1-weighted image post-contrast (**b**) show severe inflammatory changes and deformation of TMJ. **Bone marrow edema** is evident as increased signal

intensity of the condyle compared to that of the mandibular ramus marrow (* in **a**) on T2-weighted image. **Bone marrow enhancement** is shown as higher signal intensity of the condyle compared to that of the mandibular ramus marrow (* in **b**) on the T1-weighted post-contrast image. A small **joint effusion**, with high signal intensity on T2-weighted image, is seen in the upper posterior joint recess with a width exceeding 1 mm. Mild **synovial thickening** is evident as intermediate signal intensity tissue involving the anterior portions of the upper and lower joint compartments on T2-weighted image, with largest thickness less than 2 mm in the lower anterior joint recess. Both the joint fluid and thickened synovium demonstrate high signal intensity on post-contrast T1-weighted image, similar to that of veins (*arrowhead* in **b**), which indicates severe **joint enhancement** involving the entire joint. There is severe **condylar flattening** with involvement of the whole condyle. Mild **erosions** of the condyle are seen as irregular surface and thickness of subchondral bone (*arrowheads* in **a**) involving less than 50% of the condylar surface (not shown). The articular **disk** is flattened and stretched, and apparently normally located.

Figure 2: Six-year-old girl with oligoarticular JIA under methotrexate therapy. Corresponding sagittal oblique, fat-saturated T2-weighted image pre-contrast (**a**) and fat-saturated T1-weighted image post-contrast (**b**) show no active inflammatory changes but mild deformation of the condyle. **Bone marrow signal** of the condyle is normal, with the same intensity as mandibular ramus marrow (*) both on T2-weighted image and post-contrast T1-weighted image, which would be graded as absent bone marrow edema and enhancement. A minimal amount of **joint fluid** is evident in the anterior inferior joint recess with high signal intensity on T2-weighted image (*arrow* in **a**) and partly high signal intensity on post-contrast T1-weighted image (*arrow* in **b**) indicating no joint effusion and normal joint enhancement. No evidence of synovial thickening. Only the anterior circumference of the **condyle** is flattened and its surface is smooth without erosions. The articular **disk** demonstrates normal bow-tie shape and its posterior band is located normally at the apex of the condyle.

Table 1: Clinical characteristics of the cohort of 21 patients whose MRI examinations of temporomandibular joints were used for the reliability exercise. Laboratory and physical examination test values are those available at the closest date within three months in relation to the study MRI date. Abbreviations: ANA, Anti-Nuclear Antibody; DMARD, Disease Modifying Anti-Rheumatic Drug; HLA-B27, Human Leukocyte Antigen B27; JIA, Juvenile Idiopathic Arthritis; RF, Rheumatoid Factor; SD, Standard Deviation; TMJ, Temporomandibular Joint.

Table 1. Clinical Characteristics of Patient Sample (N=21)	
Mean age at diagnosis (years)	6.9 (SD 3.7, Range 0.5-13.6)
Mean age at MRI (years)	11.5 (SD 2.85, Range 6.6-16.9)
Mean disease duration (years)	4.7 (SD 4.3, Range 0-15.7)
Gender	3 Male, 18 Female
JIA Category	
Oligoarticular, persistent	6
Oligoarticular, extended	5
Polyarticular (all RF-)	7
Enthesitis related	1
Psoriatic	1
No JIA (Lyme disease related knee arthritis)	1
HLA-B27+ (total patients tested, % of patients tested)	1 (11, 9%)
ANA+ (% of patients tested)	12 (57%)
RF+ (total patients tested, % of patients tested)	1 (20, 5%)
Uveitis (% of patients tested)	7 (33%)
Facial changes (including asymmetry, decreased condylar translation, retrognathia)	12 (57%)
Crepitation	2 (10%)
Decreased mouth opening (<40mm)	5 (24%)
TMJ pain	5 (24%)
History of DMARD use (current and/or past)	13 (62%)
Active treatment with DMARDs	11 (52%)

Table 2: Inter-reader reliability scores of the three MRI scoring systems and their constituent items. AvICC (lower bound) >0.8 and SDD% <30 were the arbitrary thresholds for acceptable reliability (15). *Since these three items are only scored in the American system, a new subgroup was created excluding these items (American equivalent). This method enabled comparability between the three systems. Abbreviations: avICC, average-measure Intraclass Correlation Coefficient; CI, Confidence Interval; PCA, Percent Close Agreement; PEA, Percent Exact Agreement; SDD, Smallest Detectable Difference; sICC, single-measure Intraclass Correlation Coefficient.

Table 2: Inter-reader reliability scores of the three MRI scoring systems and their constituent items.

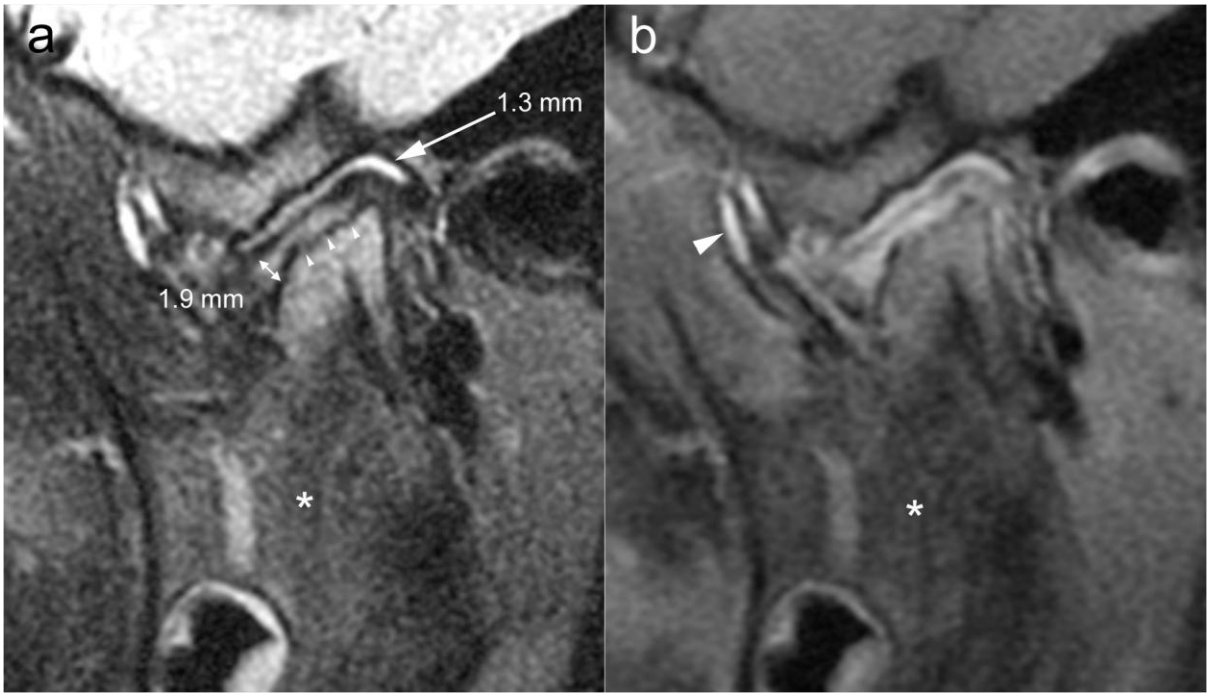
Scoring System, Domain, or Item			Coefficients of Reliability								
	Scoring System	Diagnostic Items	sICC	95% CI (lower, upper)		avICC	95% CI (lower, upper)		SDD (%)	PEA (%)	PCA (%)
Total System Scores	Swiss		0.72	0.55	0.89	0.96	0.93	1.00	22	33	69
	American		0.73	0.57	0.89	0.96	0.93	0.99	17	21	51
	German		0.64	0.45	0.82	0.96	0.93	0.99	21	21	50
Inflammatory Domain	Swiss		0.53	0.31	0.75	0.92	0.84	1.00	29	42	83
	American equivalent		0.45	0.25	0.65	0.91	0.85	0.97	25	28	66
	American complete*		0.52	0.31	0.74	0.92	0.86	0.98	24	25	61
	German		0.34	0.15	0.53	0.91	0.85	0.97	31	24	58
Osteochondral Domain	Swiss		0.76	0.61	0.91	0.96	0.93	1.00	23	61	87
	American equivalent		0.78	0.66	0.90	0.97	0.94	0.99	17	53	80
	American complete*		0.82	0.69	0.94	0.97	0.94	1.00	16	47	79
	German		0.80	0.68	0.92	0.97	0.94	0.99	18	53	83
Item-Specific Scores	Swiss	Inflammation	0.53	0.31	0.75	0.92	0.84	1.00	29	42	83
		Osseous deformity	0.76	0.61	0.91	0.96	0.93	1.00	23	61	87

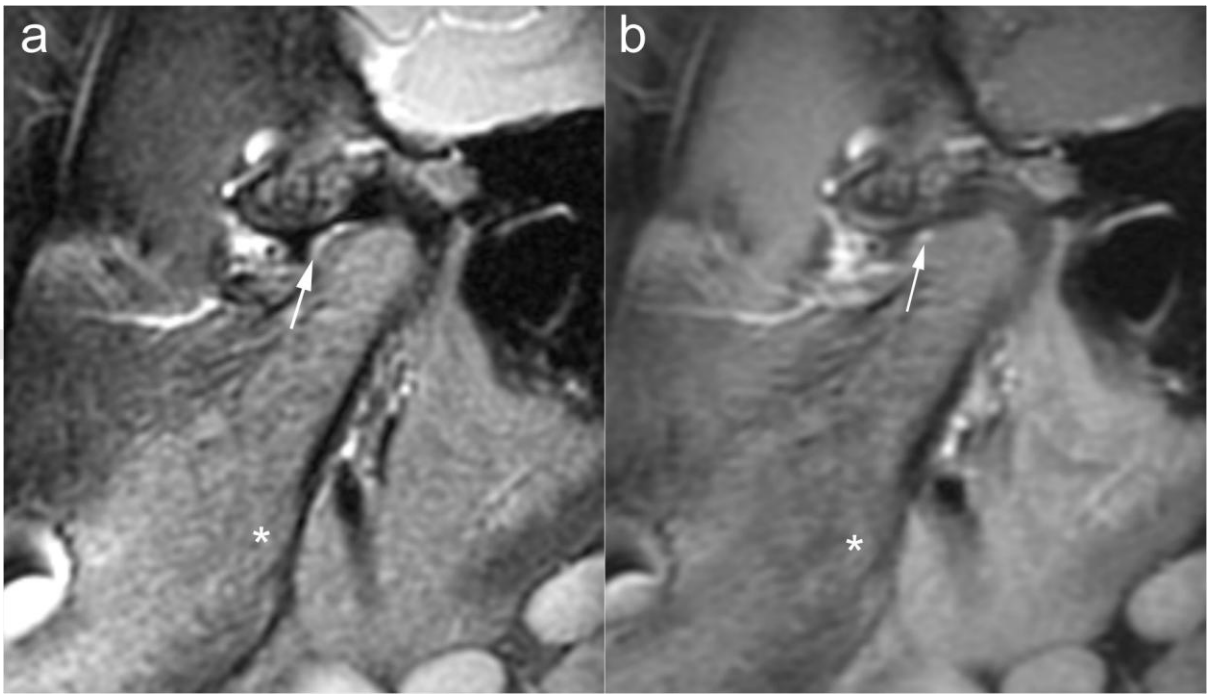
American	Marrow Edema (0-1)	0.06	0.00	0.15	0.61	0.43	0.79	63	61	100
	Effusion (0-3)	0.47	0.30	0.64	0.88	0.80	0.96	26	55	95
	Synovial Enhancement (0-3)	0.42	0.19	0.64	0.91	0.83	0.99	33	41	88
	Condyle (0-4)	0.72	0.55	0.90	0.95	0.91	1.00	23	62	90
	Disk (0-3)*	0.54	0.27	0.82	0.90	0.77	1.00	19	80	97
	Erosions (0-3)	0.69	0.47	0.91	0.94	0.83	1.00	20	82	96
	Pannus (0-3)*	0.36	0.15	0.58	0.81	0.62	1.00	30	82	89
	Osteophytes (0-1)*	0.15	0.03	0.27	0.58	0.30	0.86	35	88	100
German	Synovial fluid (0-3)	0.40	0.21	0.59	0.85	0.74	0.96	25	56	97
	Bone marrow edema (0-3)	0.01	0.00	0.11	0.57	0.34	0.80	45	54	77
	Synovitis (0-3)	0.35	0.15	0.55	0.90	0.83	0.97	33	42	89
	Erosions (0-3)	0.66	0.39	0.93	0.94	0.80	1.00	20	81	95
	Condylar shape (0-3)	0.75	0.57	0.93	0.96	0.91	1.00	28	65	91

Table 3 A,B): The consensus MRI scoring system for temporomandibular joint involvement in juvenile idiopathic arthritis. The items are grouped by A) Inflammatory and B) Damage domains. Each joint is scored independently.

	A) Inflammatory Domain				
	Bone Marrow Edema	**Bone Marrow Enhancement**	**Effusion**	**Joint Enhancement**	**Synovial Thickening**
Definition	Compared to the mandibular ramus, hyperintense marrow signaling within the condyle on T2w FS or STIR images, and/or hypointense signaling on pre-contrast T1w images without FS.	Compared to the mandibular ramus, hyperintense marrow signaling within the condyle on post-contrast T1w FS images.	Increased joint fluid with isointense signaling of joint space compared to that of cerebrospinal fluid on T2w FS or STIR images.	Signal intensity of the synovium, capsule, and joint space fluid higher than that of muscle on post contrast T1w FS images.	Thickened synovial lining of the TMJ with intermediate signal on T2w images.
Grading	Absent	Absent	Normal: ≤1mm fluid in joint recess.	Normal: High signal intensity confined to signal perimeter of normal amount of joint fluid on corresponding fluid-sensitive image	Normal: no synovium visible (joint space ≤1 mm width)
	Present	Present	Mild: >1 and ≤2mm fluid in recess or involving entire joint compartment	Mild: High signal intensity focally exceeding signal perimeter of physiologic amount of joint fluid on corresponding fluid-sensitive image	Mild: >1 and ≤2mm thickness at the point of maximum synovial thickening.
			Moderate/Severe: >2mm fluid in recess or involving entire joint compartment	Moderate/Severe: High signal intensity diffusely involving one or both joint compartments	Moderate/Severe: >2mm thickness at the point of maximum synovial thickening

	B) Damage Domain		
	Condylar Flattening	**Erosions**	**Disk Abnormalities**
Definition	Loss of the round or slightly rectangular shape of the condylar head, viewed in the sagittal-oblique plane.	Any irregularity or breaks of the bony joint surfaces leading to the loss of the smooth continuous surface of the bone, seen in both sagittal and coronal planes.	Any abnormality of the articular disk, including flattening, displacement, or destruction.
Grading	Normal round/slightly rectangular shape	No irregularities or deep breaks	Absent
	Mild: Extent of flattening involves part of the surface of the condyle	Mild: Presence of irregularities involving only part of the articular surface of the condyle	Present
	Moderate/Severe: Extent of flattening involves the entire surface of the condyle, or loss of height in the condyle	Moderate/Severe: Presence of deep breaks in the subchondral bone seen in two planes, or irregularities involving the entire articular surface of the condyle	





Appendices

Appendix 1: German scoring system (10). Each joint is scored independently, with 15 representing the most severe disease and 0 the least. *** Included instruction: "Irregular hypointense lesions (T1) with interruption of the cortical line in at least 2 planes with possible Gd enhancement (T1 fs post-Gd)."

German score (additive)				
Synovial Fluid / Effusion	*Bone Marrow Edema/Osteitis*	*Synovitis (synovial Gd enhancement and synovial hypertrophy)*	*Erosions****	*changes in condylar shape*
Grade 0 = none	Grade 0 = none	Grade 0 = none	Grade 0 = none	grade 0= no changes
Grade 1 = mild (< 2 mm) synovial fluid accumulation covering not the entire joint space	Grade 1 = mild (< 1/3 of the condyle)	Grade 1 = mild measurable (< 1 mm) synovial tissue with Gd enhancement	Grade 1 = mild (< 1/3 of the condylar surface)	grade 1=mild (<1/3 of height) flattening
Grade 2 = moderate (2-4 mm) synovial fluid accumulation with or without covering the entire joint space	Grade 2 = moderate (< 2/3 of the condyle)	Grade 2 = moderate (1-3 mm) synovial tissue with Gd enhancement	Grade 2 = moderate (< 2/3 of the condylar surface)	grade 2=moderate (<2/3 of height) flattening
Grade 3 = severe (>4 mm) synovial fluid accumulation covering the entire joint space	Grade 3 = severe (> 2/3 of the condyle)	Grade 3 = severe (> 3 mm) synovial tissue with Gd enhancement	Grade 3 = severe (> 2/3 of the condylar surface), multiple erosions covering the entire condylar surface	grade 3=severe (>2/3 of height) flattening, total loss deformation of the condyle

Appendix 2: American scoring system (11). Each joint is scored independently, for both active inflammatory and chronic changes. Active inflammatory disease score consists of items I-III (0-7), and chronic score consists of items IV-VIII (0-14).

American score (additive)

I. Marrow Edema	**II. Effusion**	**III. Synovial enhancement**	**IV. Condyle**	**V. Disk**	**VI. Erosions**	**VII. Pannus**	**VIII. Osteophyte Formation**
No edema = zero points	None = zero points	None = zero points	Normal = zero points	Normal position = zero points	None = zero points	None = zero points	None = zero points
Edema present = 1 point	Thin, diffuse = 1 point	1 mm (mild) = 1 point	1/3 flattened = 1 point	1/3 destruction = 1 point	In 1/3 = 1 point	≤ 1 mm = 1 point	1 or more (positive) = 1 point
	Focal effusion >1mm= 2 points	2 mm (moderate) = 2 points	2/3 flattened = 2 points	2/3 destruction = 2 points	In 2/3 = 2 points	1 – 2 mm = 2 points	
	Entire space >1mm = 3 points	≥ 3 mm (severe) = 3 points	Complete flattened = 3 points	Complete destruction = 3 points	Complete destruction = 3 points	In 3/3 (complete) = 3 points	
			Condyle enlarged = 4 points				

Appendix 3: Swiss scoring system (12). Each joint is scored independently. Both the inflammation and osseous deformity domain are scored progressively, by the presence of the most severe change.

Swiss grading system (progressive)

	Inflammation	**Osseous deformity**	
0	≤ small amount of joint fluid (high signal T2w), - enhancement confined to joint fluid	- deep s-shaped mandibular fossa, - round (young patient) / squared condyle (older patient)	**0**
1	- mild **joint space enhancement exceeding joint fluid** (T2w), ± bone marrow oedema / enhancement	- **mild flattening** of condyle - mild flattening of mandibular fossa / eminentia	**1**
2	- intense joint **enhancement involving entire joint space**, ± enhancing joint effusion ± bone marrow oedema	- moderate flattening of condyle - **moderate flattening** of mandibular fossa / eminentia	**2**
3	grade 2 **+ synovial thickening** (intermediate signal T2w)	- **severe flattening** of condyle with loss of height - flat mandibular fossa / eminentia **± small erosions**	**3**
4	**joint space expanded** by enhancing soft tissue ± bone marrow oedema ± intraaarticular calcification / ossification	- **fragmentation** of condyle **± large erosions** of condyle or eminentia **± bone apposition** in mandibular fossa / on condyle	**4**

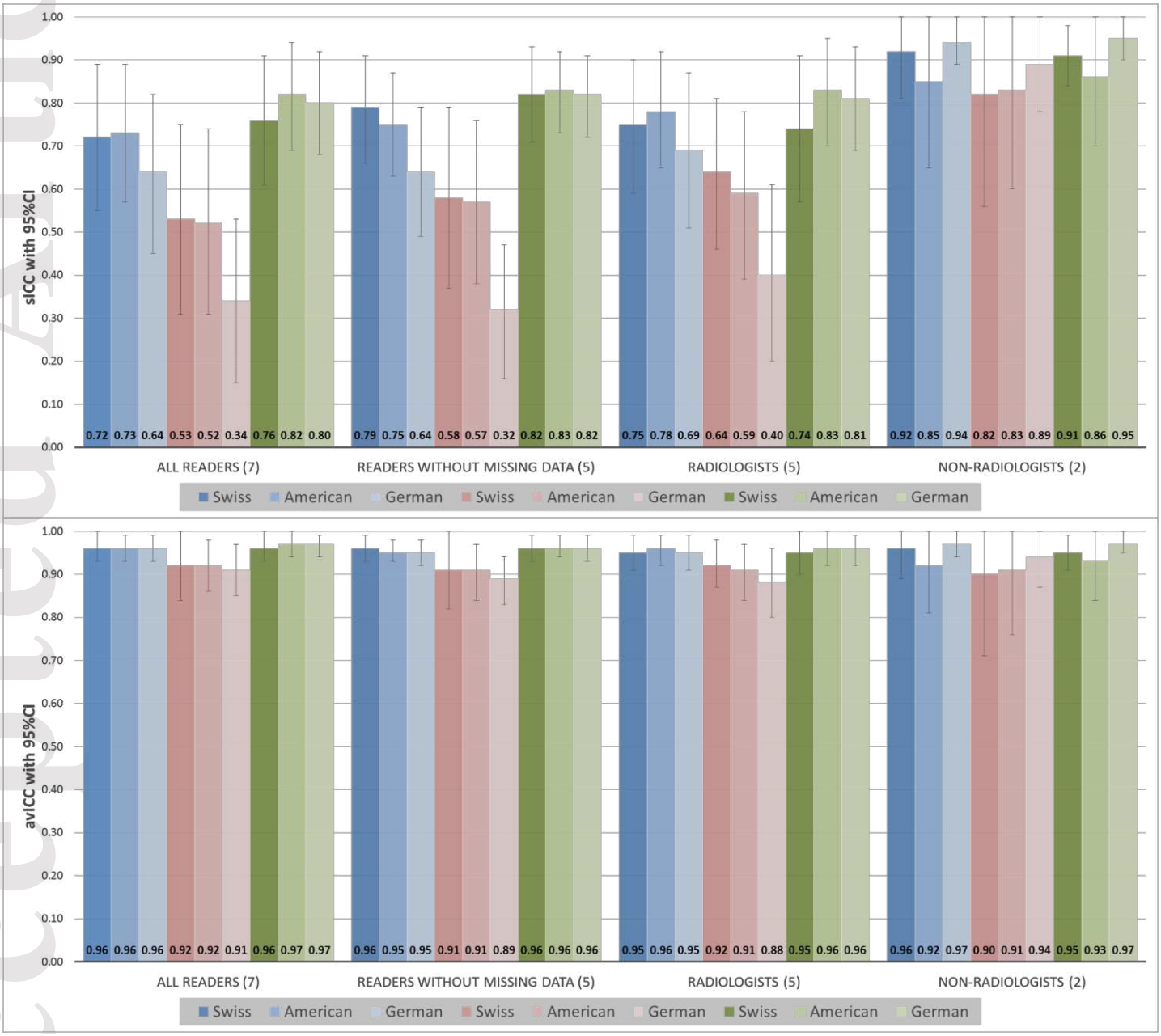
Appendix 4: Affiliating institutions of the expert group involved in the scoring system development. The table lists the affiliating institution, country, and speciality of the collaborators involved in the development of the consensus TMJ MRI scoring system.

Name	Specialty	Institution	City, Country	Reliability Exercise Participation	Consensus Meetings
Andrea S. Doria	Pediatric Radiologist	Hospital for Sick Children	Toronto, Canada		✓
Bernd Koos	Orthodontist	University Hospital Tübingen	Tübingen, Germany	✓	✓
Christian J. Kellenberger	Pediatric Radiologist	University Children's Hospital Zürich	Zürich, Switzerland	✓	✓
Elka Miller	Pediatric Radiologist	Children's Hospital of Eastern Ontario	Ottawa, Canada	✓	✓
Emilio J. Inarejos	Pediatric Radiologist	Hospital Sant Joan de Deu	Barcelona, Spain		✓
Jennifer Stimec	Pediatric Radiologist	The Hospital for Sick Children	Toronto, Canada	✓	✓
Lynn Spiegel	Pediatric Rheumatologist	The Hospital for Sick Children	Toronto, Canada		✓
Marinka Twilt	Pediatric Rheumatologist	Alberta Children's Hospital	Calgary, Canada		✓
Marion A. van Rossum	Pediatric Rheumatologist	Academic Medical Centre	Amsterdam, The Netherlands		✓
Nikolay Tzaribachev	Pediatric Rheumatologist	Pediatric Rheumatology Research Institute	Bad Brahmstedt, Germany	✓	✓
Randy Q. Cron	Pediatric Rheumatologist	Children's Hospital of Alabama	Birmingham, Alabama, USA		✓
Rotraud K. Saurenmann	Pediatric Rheumatologist	University Children's Hospital Zürich	Zürich, Switzerland		✓
Saurabh Guleria	Pediatric Radiologist	Austin Radiological Association	Austin, Texas, USA	✓	✓
Tal Laor	Pediatric Radiologist	Cincinnati Children's Hospital Medical Center	Cincinnati, Ohio, USA		✓
Thekla von Kalle	Pediatric Radiologist	Olgahospital Klinikum Stuttgart	Stuttgart, Germany	✓	✓
Tore A. Larheim	Maxillofacial Radiologist	University of Oslo	Oslo, Norway		✓
Troels Herlin	Pediatric Rheumatologist	Aarhus University Hospital	Aarhus, Denmark		✓
Yoginder Vaid	Pediatric Radiologist	Children's Hospital of Alabama	Birmingham, Alabama, USA		✓

Appendix 5: MRI protocol for the images used in the reliability exercise. Abbreviations: FOV, field-of-view; FS, fat suppression sequence; FSE, fast spin echo; FSPGR, fast spoiled gradient recalled echo; +Gd, post gadolinium injection; mm, millimeter; PD, proton density weighted sequence; SE, spin echo, TE, echo time; TR, repetition time.

Imaging Sequence (in order of acquisition from left to right)	T1 FSPGR	PD FSE	T2 FSE FS	T1 FSE FS +Gd	T1 SE FS +Gd	3D FSPGR +Gd
Plane	Sagittal oblique	Sagittal oblique	Sagittal oblique	Sagittal oblique	Coronal	Sagittal oblique
TE	4.2	25	86	11	19	10.4
TR (ms)	325	2660	2840	600	600	4.2
Flip angle	80	90	90	90	90	20
FOV (mm x mm)	120	120	120	120	160	100
Acquisition Matrix	384 x 224	256 x 224	256 x 224	256 x 224	256 x 192	256 x 192
Slice thickness (mm)	2	2	2	2	2	2
Slice spacing (mm)	2	2	2	2	2	1
Echo train length	-	8	16	3	-	-

Appendix 6: Sensitivity analysis clustering various groups of readers. Top) single measure intraclass correlation (sICC) scores, bottom) average measure intraclass correlation (avICC) scores. 95% confidence intervals (95%CI) of the ICCs are obtained from 500 bootstrap replications. ICCs across groups of readers may not be comparable, since the number of readers affect score magnitude. ICCs were recalculated without the two readers who had missing data to test for potential missing data bias. Missing data did not affect the results. Non-radiologists (n=2) showed higher inter-reader reliability in sICC than radiologists, though when adjusted for the number.



Total Score Inflammation Sub-score Damage Sub-score

Appendix 7: Structured and unstructured development phases of the TMJ MRI scoring system. Two rounds of Delphi sessions and a two-stage nominal group technique were conducted to formalize the consensus scale development process. Reliability results from a case reading exercise using three existing scoring systems were used to inform item selection, definition, and grading. Abbreviations: NGT, nominal group technique.